Chapter 5

Quantitative Evaluation of the Linkage Operations of the 1996 Census Reverse Record Check

Julie Bernier, Statistics Canada

Abstract

A probabilistic linkage of two files is performed using the theory derived by Fellegi and Sunter (1969). The decision on whether a unit from each file are linked is based on the linkage weight obtained. In effect the linkage weight, a one-dimensional variable, is divided into three ranges: one for which links are accepted, one for which they are rejected and the intermediate range, where a link is possible. Manual inspection of possible links is needed to decide which ones represent the same unit. At the end of the linking procedure, the accepted links and those possible links that were confirmed by the manual check are kept. Under certain conditions, the results of this check provide all the information needed for a quantitative evaluation of the quality of the links. In this article we present a brief description of the Reverse Record Check (RRC) and the role of probabilistic linkage in this project. We then offer a definition of the reliability of a link and describe a procedure for estimating the minimum reliability of the links kept, using a fitted logistic regression model based on the manual checking of the possible links. Finally, we present the results obtained for the RRC96, describing the number of links obtained and the reliability of those links.

Introduction

When a probabilistic linkage is performed between two files, any of several approaches may be used. Depending on the approach chosen, it may be that among the linkage weights obtained, there will be a limited number of different values. In this case, a number of links are associated with each weight, and we can proceed by sampling to estimate the proportion of true links for each possible weight value. The next step is to manually inspect the links sampled. It may also happen that the set of possible values of the linkage weight will be quite varied. This may result in the use of a great number of comparison rules, each making a different contribution to the total weight depending on whether or not there is a match between the fields compared. This variety of weights may also result from the use of comparison rules that assign frequency weights in the event of a match. The use of frequency weights means that where there is a match, a different contribution is made to the total weight depending on whether the content of the fields compared is more or less frequent in the population. For example, a larger contribution is made when there is a match on a relatively rare family name. In the case of a set of varied weights, the distribution of links on the basis of the linkage weights closely resembles a continuous distribution. The proportion of true links may then be estimated by grouping the weights by intervals or by using a logistic regression. The use of logistic regression was chosen as the method of estimating the proportion of true links in the linkage of the 1996 Reverse Record Check (RRC96) with the 1990 Revenue Canada files (RCT90), since in that linkage, a number of comparison rules were involved. Furthermore, for two of the fields compared, namely family name and the first three entries of the postal code, frequency weights were used.

The Reverse Record Check

The purpose of the reverse record check is to estimate the errors in coverage of the population and of private households in the Census. It also seeks to analyse the characteristics of persons who either were not enumerated or were enumerated more than once. The method used is as follows:

- Using a sample frame independent of the 1996 Census, a sample is drawn of persons who should have been enumerated in the Census.
- A file is created containing as much information as possible on these persons and their census families.
- If possible, the addresses of the selected persons (SP) and their family members (close relatives living under the same roof) are updated using administrative files.
- Retrieval operations are carried out by interviewers in order to contact the selected person and administer a questionnaire to him or her. The purpose of the questionnaire is to determine the addresses at which the person could have been enumerated.
- Search operations are carried out on the questionnaires and in the Census database in order to determine how many times the selected person was enumerated.

The Role of Probabilistic Linkage

Probabilistic linkage is used in the address updating procedure. In this procedure there are two principal stages. First, probabilistic linkage of the RRC96 with the Revenue Canada 1990 (RCT) file is carried out. The reason for choosing the year 1990 is that this database was created in early 1991 and the sample frame of the RRC is largely made up of the database of the 1991 Census and the files of the RRC91. When this linkage is successfully completed, we obtain the social insurance number (SIN) of the selected person or a member of that person's family. In the second stage, an exact linkage is made between the RRC96 and the 1991, 1992, 1993, and 1994 Revenue Canada files in order to obtain the most recent address available in those files. For this linkage, the SIN is used as an identifer. It is by means of these addresses that we can begin tracing the selected persons by the RRC.

During operations to link the RRC sample with the 1990 Revenue Canada files, we determined, for each of the eight region-by-sex groups, a threshold linkage weight beyond which all links were considered definite or possible and were retained for the next stage. Subsequently, we checked the weakest links in order to determine whether they were valid or false. This enabled us firstly to eliminate the false links before proceeding to subsequent operations and secondly to determine the reliability of the links retained. Two other approaches may be used. One can define a fairly low linkage weight beyond which all links are kept without being checked. This yields a greater number of links, some of which have little likelihood of being valid. There are two drawbacks to this approach. First, it means that the interviewers responsible for tracing selected persons are given more false leads. This can result in time loss during tracing and a greater probability of interviewing by error a person other than the one selected in the sample. Second, the update address is processed in the search operation. This too can needlessly increase the size of this operation. The other possible approach is to define a fairly high linkage weight beyond which all links are retained without being checked. They yields fewer links, but those obtained have a strong probability of being valid. The disadvantage of this method is that it increases the number of persons not traced. This type of nonresponse is more common in the case of persons living alone, and such persons are also the ones who have the greatest likelihood of not being enumerated. It is for this reason that we preferred the approach that requires manual checking but serves to reduce this type of nonresponse without needlessly expanding the tracing and search stages.

Checking Procedure

In light of the amount of data to be processed, linkage is carried out separately in eight groups defined by the sex and the geographic region of the individual to be linked. The four geographic regions are: Eastern Canada, Quebec, Ontario, and lastly, Western Canada and the Northwest Territories. For each of the four regions it is necessary to define a grey area within which links are considered "possible" rather than being accepted automatically. This area extends from the lower boundary weight (LOW) to a weight determined iteratively (UPP) in the course of checking. The point LOW is determined through guesswork, by examining the links ranked by descending order of weight. The persons engaged in this task try to choose a point LOW such that manual checking will be done when links have a fairly high probability of being valid (approximately 75%). The checking begins with UPP chosen such that the grey area contains roughly 1.5% of the links retained for the region in question. To reduce the workload, some of these links are then checked automatically, in the following manner: when both spouses in a household have linked, if one of the two (C1) obtained a high linkage weight and if in addition that person's record at RCT is found to contain the SIN of the spouse (C2), and if that SIN is the same as the one found in the record that is linked with C2, then the link of C2 with RCT is considered reliable, even if it obtained a linkage weight within the grey area. All the links in the grey area that did not satisfy the foregoing criterion were checked manually. These checks were carried out using all available information on the household as a whole. After the entire grey area was checked, if the number of rejected links seemed high, UPP was changed so as to add from 1.5% to 2% more links to the grey area. These two steps (choosing UPP and checking) were repeated until the rejection rate for the links checked seemed lower than 10% for links with a linkage weight close to UPP.

Shown below, for each region, are the grey area boundaries, the percentage of links within those boundaries and the total percentage of links rejected in the grey area.

Table 1. -- Results of Checking

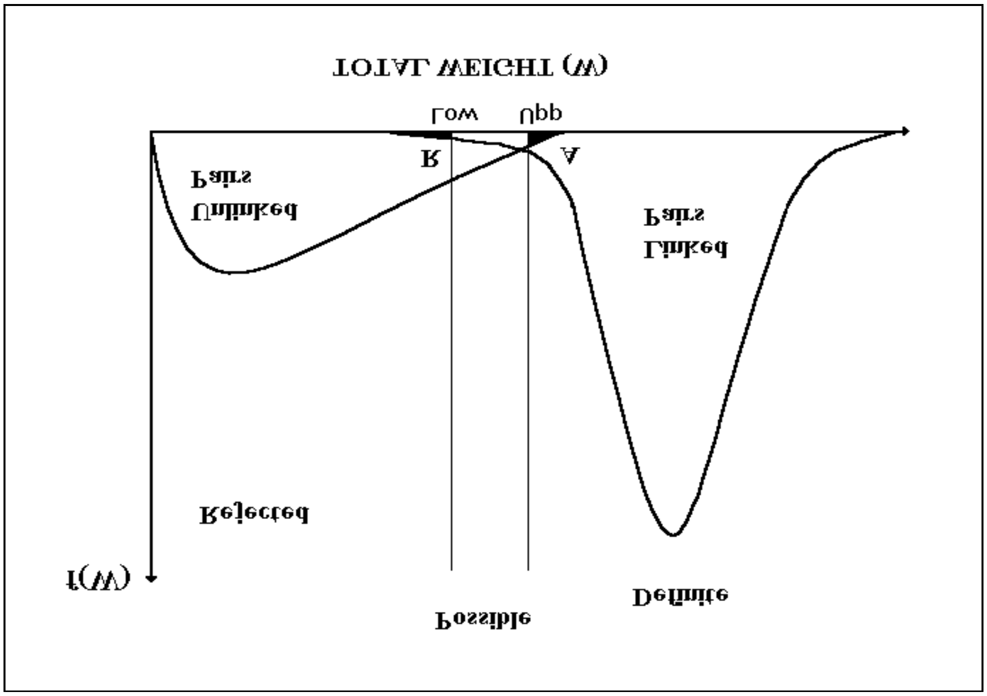
Region	LOW	UPP	Percentage of Links Checked	Percentage of Links in Grey Area Rejected
Eastern	222	244	1.5	2.1
Quebec	221	304	7.5	23.0
Ontario	274	309	2.0	1.2
Western	219	258	1.5	4.9

As stated in the introduction, these checks are useful in two ways. First, they serve to eliminate most of the false links. They therefore enhance the quality of the set of links obtained. Second, these checks enable us to form a data set that contains, for various links, both their status (valid or false) and their linkage weight. Using this data set, we were then able to assess the reliability of the accepted links.

Definition of the Reliability of a Link

The probabilistic linkage procedure consists in calculating, for each pair of records, a weight W based on whether the fields compared match or do not match and on the probability of matching these fields given a linked pair or an unlinked pair. Generally, during the matching procedure, we try to determine a lower boundary and an upper boundary such that pairs with a weight lesser than LOW are rejected, those with a weight greater than UPP are accepted and those between these two boundaries are considered as possible links and eventually undergo another classification procedure. The following figure illustrates these concepts.

Figure 1.--Distributions of Pairs of Records by Linkage Weight



In linkage, two types of error are possible: accepting a pair that is not linked (A) or rejecting a linked pair (R). We are usually interested in the following probabilities:

P(accept a link | the pair is not linked) = P(W>UPP | the pair is not linked) and
P(reject a link | the pair is linked) = P(W<LOW | the pair is linked),

which are called ***classification error probabilities***. We try, then, to choose LOW and UPP such that these two probabilities meet certain optimization criteria (see Fellegi and Sunter). Methods for estimating these probabilities may also be obtained by using samples of accepted links and rejected links that are checked manually (see Armstrong and Mayda, 1993 for a partial review of these methods). For the RRC96, we proceed differently. We determine a point LOW below which all links are rejected, but we do not define in advance a point UPP that would separate possible links from definite links. This point is instead determined during the manual check when it is felt that the links checked exhibit a high enough frequency to stop checking. Here we are instead interested in the following probabilities:

P(the link is valid | the link is accepted) = P(valid | W>UPP) and
P(the link is valid | the link is rejected) = P(valid | W<LOW).

These two probabilities will be called ***the reliability of accepted links and the reliability of rejected links.*** It should be noted that the term reliability applies here to a link and not to the procedure that leads to the acceptance or rejection of this link. We therefore speak of the reliability of a rejected link as being the probability that this link is valid, which in fact amounts to a classification error. We could estimate these two probabilities respectively by the proportion of linked pairs among the accepted links and the proportion of linked pairs among the rejected links. These estimates would require manual checking of two samples drawn respectively from the accepted links and the rejected links. We ruled out this method for two reasons. First, the rejected links set was not retained. Second, for an estimate of a very low error rate to be acceptable, a very large sample is required, which means that the more successful the linkage procedure, the more costly the quantitative evaluation of the reliability of links using two samples. We therefore chose an alternative that allows us to use checking in the grey area rather than requiring checking of one or two additional samples.

Reliability Evaluation Procedure

We can speak generally of P(valid | W>UPP), which is the reliability of the links in the accepted links set, and of P(valid | W<LOW), the reliability of the links in the rejected links set; but we cannot speak more specifically of P(valid | W), the proportion of valid links in the subset consisting of pairs with linkage weight W. The proportion P(valid | W) may be defined as the reliability of a link of weight W. When we speak of quantitative evaluation, we may want to obtain a general estimate of the reliability of the accepted links and the rejected links, or we may want more specifically to estimate P(valid | W) for certain critical values of W. Since this probability increases with W, we have only:

P(valid | W=UPP) constitutes a lower boundary for the reliability of the accepted links.
P(valid | W=LOW) constitutes an upper boundary for the reliability of the rejected links.

In addition, no error is associated with the grey area, meaning that we consider that the manual check is a total success. Our quantitative evaluation therefore consists in estimating P(valid | W=UPP) and P(valid | W=LOW). To do this we use a logistic regression model, the parameters of which are estimated from the links in the grey area. This method is based on two assumptions. First, it must be assumed that the variable W is linked linearly to the logit function of the reliability to be estimated (logit(p)=log(p/1-p)). The logistic model is of the following form:

logit(p | W) = a + b W, where p is the probability that the link is valid.

This condition, which constitutes a test of goodness of fit for the model, is verified by a method described in the appendix. Second, the grey area must contain a sufficient number of unlinked pairs with various W values. When the number of unlinked pairs in the grey area is insufficient, the hypothesis $\beta=0$ cannot be rejected at a meaningful level. In the latter case, the proportion of valid links in the grey area is very high and can serve as the upper boundary for P(valid | LOW) and the lower boundary for P(valid | UPP). It should be noted that such a situation means that we have been too strict in choosing the cutoff point LOW in the linkage operations, and have therefore rejected many valid links and inappropriately used manual checking on a set of links with very high reliability. The procedure proposed is therefore the following:

- Check links in the grey area; each pair is considered linked or unlinked.

- Estimate parameters α and β of the logistic regression.

- Test the goodness of fit of the model and test the hypothesis $\beta=0$.

- If the results of the tests are satisfactory, estimate P(valid | W=UPP) by using
  logit(P(valid | W=UPP)) = α +β UPP and estimate P(valid | W=LOW) by using
  logit(P(valid | W=LOW)) = α +β LOW.

- If the results of the tests do not allow us to use the model, we merely estimate the proportion of valid links in the grey area.

Results

Shown below are the results obtained for the four regions. The estimates are made using both males and females, since introducing the sex variable into the logistic regression does not make a significant contribution.

Table 2. -- Estimate of Reliability

Region	**Eastern**	**Quebec**	**Ontario**	**Western**
Estimate of regression equation		logit(p)= -2.82 + 0.0165 W		logit(p)= -12.70 +0.0665 W
Estimate of reliability at W=LOW		69.6%		86.6%
Estimate of reliability at W=UPP	> 97.9%	90.0%	> 98.8%	98.8%
Estimated W for which reliability is 90%		304		224

For Eastern and Ontario regions, we didn’t find enough unlinked pairs to do a logistic regression. This means that we could probably have set LOW lower in the linkage operations. For Quebec and Western regions, we estimated the reliability using the logistic regression model. It will be recalled that we check either 1.5% of the weakest links or several series of links until the estimated reliability at W=UPP seems to us to be greater than 90%. For Region 2, the estimate of 90% for reliability at UPP shows that we succeeded in choosing UPP such as to ensure good reliability of links while minimizing manual checking.

It should be recalled that:

- All links in the grey area were checked, and those that were false were rejected.

- The estimated reliability at point UPP is a lower boundary for the reliability of the accepted links.
- The overall reliability in the interval [LOW,UPP] is also a lower boundary for the reliability of the accepted links.

We therefore estimate that the accepted links have a reliability greater than 97.9% in the Eastern region, greater than 90% in Quebec, greater than 98.8% in Ontario, and greater than 98.8% in the Western region.

It should lastly be noted that often in linkage procedures, the approach used is one that seeks to retain as many links as possible. In such cases, the LOW and UPP boundaries are set much less strictly than was done for the RRC-RCT linkage. In that situation, using the method described here could prove to be ineffective or even discouraging, since the reliability calculated by means of logistic regression is a lower boundary for the reliability of the accepted links. In some cases, that boundary could be very low although the overall rate of false links is acceptable. In such cases, it may be preferable to instead use a sample of the accepted links to estimate reliability generally.

Linkage Results and Conclusion

After choosing LOW and UPP and determining the links to be retained (either automatically or by manual checking), we obtain the following linkage rates for Canada's different geographic regions:

Table 3. -- Linkage Results by Region

Region	Eastern	Quebec	Ontario	Western
Selected person(SP) linked	57%	54%	58%	54%
SP not linked but other family member linked	36%	35%	31%	36%
No linkage	6%	10%	9%	9%
Linkage not attempted	1%	1%	2%	1%
Sample size	12,440	7,328	9,243	16,820

As may be seen, an update address is obtained for more than half of the selected persons, with an address reliability greater than 90%. As regards persons who are not linked, in many cases another member of the household is linked, so that we can nevertheless obtain a valid address for tracing in roughly an additional 35% of cases.

These results should enable us to obtain a satisfactory response rate for the RCC96.

To verify the linearity of the relationship between the logit of reliability and weights W, we grouped the weights into intervals and worked with the midpoints of these intervals. For the two regions where a model has been used, the model obtained in this way is very close to the one obtained by means of logistic regression. This confirms that the logistic model functions well for predicting the reliability of the links in the manual checking range. This model could also be used on a sample of links checked during the linkage procedure, so as to determine UPP and LOW points that result in both an acceptable level of reliability and a reasonable amount of manual checking, or even to choose to change the linkage rules if we suspect that it will not be possible to achieve these two objectives simultaneously.

References

Armstrong, J. B. and Mayda, J. E. (1993). Model-Based Estimation of Record Linkage Error Rates, *Survey Methodology*, 19, 2, 137-147.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.